Context and Copying in Neural Machine Translation

Context and Copying in Neural Machine Translation

Rebecca Knowles
Dept. of Computer Science
Johns Hopkins University
rknowles@jhu.edu

Philipp Koehn
Dept. of Computer Science
Johns Hopkins University
phi@jhu.edu

Abstract

Neural machine translation systems with subword vocabularies are capable of translating or copying unknown words. In this work, we show that they learn to copy words based on both the context in which the words appear as well as features of the words themselves. In contexts that are particularly copy-prone, they even copy words that they have already learned they should translate. We examine the influence of context and subword features on this and other types of copying behavior.

1 Introduction

In translation, certain tokens – often names and numbers – should be copied from the source sentence to the target sentence. Word copying is fairly straightforward in phrase-based statistical machine translation, where unknown words can be left untranslated (copied to the target). It poses more of a challenge in neural machine translation systems, which often use limited or subword vocabularies and soft attention rather than strict alignment. This has resulted in a variety of approaches to copying, which make use of pre-/post-processing and/or network modifications (e.g. explicit switching between generation and copying).

Neural machine translation models that use subword vocabularies to perform open-vocabulary translation have been observed to correctly translate unknown words or copy words (one subword at a time, if need be) even when the full word to be translated or copied was not observed in training. Koehn and Knowles (2017) found that neural machine translation systems using subword vocabularies outperformed phrase-based statistical machine translation systems on the translation of unknown words. This raises the questions that we seek to answer: to what extent does byte-pair encoding[1] solve the copying problem (without requiring modifications to the network structure)? More generally, what are subword neural machine translation models learning about copying?

We find that neural machine translation systems (with attention, trained on subword vocabularies) learn to copy words (both novel and observed) based on their sentential contexts. Additionally, though the models have no knowledge about the components of each subword unit, they learn that certain categories of tokens (e.g. capitalized tokens) tend to be copied. We use quantitative and qualitative evaluations to shed light on what these models learn about copying tokens and about the contexts in which copying occurs.

2 Related Work

Prior work on copying in neural machine translation has typically focused on rare or unknown words. Luong et al. (2015) augment data with word alignments to train a neural machine translation system (without attention) that emits both a translation and source word positions for any out-of-vocabulary (OOV) tokens emitted. They post-process OOVs with a dictionary or by copying. Currey et al. (2017) augment training data with monolingual target language text as bitext and find that it improves copying in low-resource settings. Ott et al. (2018) and Khayrallah and Koehn (2018) examine negative effects of source copying.

Both Gu et al. (2016) and Gulcehre et al. (2016) modify neural sequence to sequence models to explicitly perform copying. Gu et al. (2016) focus on monolingual tasks (dialogue systems and summarization), proposing a model that can both generate and copy text. Gulcehre et al. (2016) perform experiments on neural machine translation (with attention), using whole-word vocabularies (and an UNK token to represent unknown words). Their model incorporates a switching variable that determines whether to copy or generate a translation.

In this work, we focus on subword vocabularies for neural machine translation, using byte-pair en-

[1] A type of subword vocabulary (Sennrich et al., 2016b).

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 3034–3041
Brussels, Belgium, October 31 - November 4, 2018.

Data	% Tokens Copied	
	DE	*EN*
Europarl	1.8%	2.0%
News Commentary	2.9%	3.3%
Full Training (EN–DE)	7.6%	8.1%
Full Training (DE–EN)	8.6%	9.2%

Table 1: Percentage of tokens which should be copied, across training data sources.

coding (BPE, Sennrich et al. (2016b)). The other approaches described are somewhat orthogonal to the use of subword vocabularies, but may require modifications to handle subwords.

3 Data and Models

We train German–English (DE–EN) and English–German (EN–DE) neural machine translation models with attention, similar to the University of Edinburgh's WMT 2016 submissions (Sennrich et al., 2016a). Models are trained using the Marian toolkit (Junczys-Dowmunt et al., 2018).[2] We use the WMT parallel text[3] (Europarl, News Commentary, and CommonCrawl) along with synthetic backtranslated data.[4]

4 Initial Analysis

We analyze the training data to learn about the prevalence and characteristics of words that should be copied in translation and the contexts in which they occur. We consider both the full training data (including backtranslations and CommonCrawl) and cleaner subsets. We restrict our search for copied words to tokens of length 3 or more characters.[5] Our heuristic for detecting copied tokens is this: a word is a "copied token" if it appears the same number of times in both the source and target sentence.[6] As we will show, copied words tend to belong to specific categories (proper nouns, numbers, etc.) which coincide with their repeated appearance in certain contexts (e.g. names following titles like "Ms" or "Prime Minister").

[2]We use recommended settings and early stopping, with results comparable to WMT 2016 systems, with BLEU scores of 39.9 (DE–EN) and 33.2 (EN–DE) on the 2016 test set.

[3]http://www.statmt.org/wmt16/translation-task.html

[4]http://data.statmt.org/rsennrich/wmt16_backtranslations/

[5]This has the benefit of removing words like *in* which are the same in German and English, but may nonetheless be considered translations rather than copies.

[6]In DE–EN, we find one notable exception to this heuristic – *was* – which is a homograph, not a copy. It makes up $< 1\%$ of copied tokens in Europarl/News Commentary.

4.1 Where do copied words appear?

In Table 1, we see that between 1.8% and 9.2% of tokens are copied.[7] Though the majority (or near-majority) of sentences do not contain any copied words (of length 3 or more), copied words are still quite prevalent: approximately 18% of sentences in each full training dataset contain one, 4% to 5% contain four, and there is a long tail (one sentence contains 70). Sentences with many copied words often contain direct quotations, third language text (not source/target), or a sequence of copied words (e.g. comma-separated numbers or names).

The cleaner Europarl and News Commentary corpora have lower percentages of copied tokens than the overall training data. Of particular note, the backtranslated data contains some examples of copying that we'd prefer for the system *not* to learn, such as target language words appearing untranslated in the (backtranslated) source side data.

4.2 What words are copied?

We first examine the part-of-speech (POS) tags[8] of copied words. In the EN–DE training data, most copied words are tagged on the English side as NNP (proper noun, singular), including names of individuals, places, or organizations (eg. González, Wales, Union). The next most frequent categories are CD (cardinal number) – including numbers like *42* that should be copied and ones like *seven* which should be translated – and NN (noun, singular or mass). The results are similar for DE–EN training data (tagged on German with a different tag set): PROPN (proper noun) is the most frequent tag for copied words, followed by NUM (numbers) and NOUN. Punctuation would rank highly if we included short tokens.

5 Experiments and Analysis

We address two main questions: (1) Do certain *contexts* encourage copying? (2) Do certain *words* exhibit features that make them more likely to be copied (regardless of context)?

5.1 Contexts

Working from the intuition that certain contexts indicate that copying should occur – for example, a name following a title like "Ms" or "Frau"

[7]The two full training sets differ due to the synthetic backtranslated data; the rest of the corpora are identical.

[8]POS tags are generated by the Stanford POS tagger (Toutanova et al., 2003). For English: english-left3words-distsim.tagger. For German: german-ud.tagger.

S: Therefore, Mrs Ashton, your role in this is invaluable.
R: Darum, Frau Ashton, ist Ihre Aufgabe in diesem Zusammenhang von unschätzbarem Wert.
T: Therefore, Mrs [NNP], your role in this is invaluable.
E1: Therefore, Mrs BBC, your role in this is invaluable.
D1: Deshalb, Frau BBC, ist Ihre Rolle hierbei von [...]
E2: Therefore, Mrs June, your role in this is invaluable.
D2: Deshalb, Frau June, ist Ihre Rolle dabei von [...]
E3: Therefore, Mrs Lutreo, your role in this is invaluable.
D3: Daher, Frau Lutreo, ist Ihre Rolle hierbei von [...]

Table 2: Source, reference, template, and examples of template-token combinations. *E1* has a word usually (76.0% of the time) copied in training, *E2* has one rarely (0.8% of the time) copied, and *E3* has a novel one. In training, 84.8% of NNPs with this left bigram context (", Mrs") were copied.

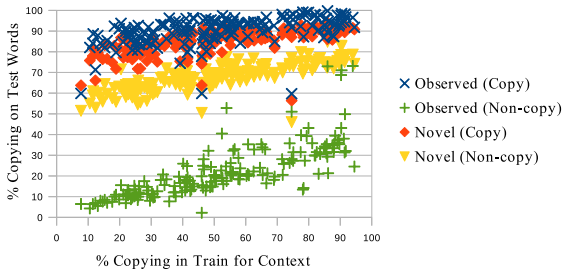

Figure 1: Percent of NNP (EN–DE) tokens copied by how copy-prone the context is, by category. Each point is the percentage of copying for all within-category words, across all example templates for one particular context (averaged over between 1,100 (novel-non-copy) and 13,150 (observed-non-copy) binary copy values).

should often be copied – we examine the relationship between context and copying, focusing on left bigram contexts. We show that the machine translation system learns that certain contexts are so indicative of copying that it will even copy (not translate) words that it has *learned to translate* if they are seen in a sufficiently copy-prone context.

For each POS, we collect a set of left bigram contexts that precede a word with that tag. We filter by frequency and diversity of tokens following the bigram.[9] For each context-POS pair, we select 50 random templates from the training data containing the bigram context followed by a word with that POS.[10] Each context-POS pair is associated with a percentage that represents how often it exhibited copying in the training data. For example, in the copy-prone context "thank Mrs [NNP]" the NNP was copied 91.1% of the time, compared to 15.3% of the time in "Republic of [NNP]".[11]

We take all word types with a given POS tag from the WMT 2016 test set, dividing them into four categories based on two binary distinctions: *observed* (in training data) or *novel* (not observed in training), and *copy* (typically copied) or *non-copy* (not typically copied) and filter the observed ones based on training frequency. We count words as *non-copy* if they were copied $\leq 30\%$ of the time, and as *copy* if they were copied $\geq 70\%$ of the time.[12] We then combine each word with each POS-appropriate example template and perform preprocessing (including BPE) and translation.[13] Table 2 shows examples.

For each context, we calculate the percentage (across all example templates for that context and all words, separated by observed/novel and copy/non-copy categories) of the time that the words in that context were copied. We then compare it to the percentage of the time that copying occurred for that context-POS tag pair in training.

Figure 1 shows NNP (EN–DE) results. Both observed-copy and novel-copy words behave almost identically, with copying percentages generally above 80%, and a slight trend upward as contexts become more copy-prone (moving to the right along the horizontal axis). Novel-non-copy words shadow these, but with a drop in copying percentage (see Section 5.3). Most interesting is the observed-non-copy category. In contexts that are not copy-prone, minimal copying occurs.[14] However, as they are placed in increasingly copy-prone contexts, even these words that the system has learned it should *translate* are being copied. We observe the same trend for words tagged NN and CD, and for PROPN, NOUN, and NUM words in the DE–EN direction. This demonstrates that the machine translation system has learned that certain contexts are copy-prone.

We manually analyze outliers that appear much

[9] See Appendix A for details and examples of contexts.

[10] We select contexts and templates from the full training data, rather than only Europarl/News Commentary, because we are interested in what patterns the model is learning from *all* data to which it has been exposed.

[11] For DE–EN translation, we see similar patterns: two of the three most copy-prone PROPN left bigram contexts are "sagte Frau" and "sagte Herr" ("said Ms/Mr"), while many less copy-prone ones end with articles.

[12] See Appendix B for details.

[13] We use the Marian batch decoder, with recommended settings: beam size 6 and length normalization penalty of 0.6.

[14] Note that some of the non-copy words were sometimes copied in training data, even if only in backtranslations.

	Drop	Change	Other
Novel-Copy	24	102	60
Novel-Non-Copy	14	128	50
Observed-Copy	51	6	126
Observed-Non-Copy	12	1	186

Table 3: Counts of automatically detected output categories (*drop*, *change*, and *other*) for a sample of NNP tokens (EN–DE) that were not copied.

more or less copy-prone than expected. In both cases, the cause appears the same: the context occurred repeatedly in many very similar sentences in the training data. Highly copy-prone contexts that produced copying percentages greater than 70% even in observed-non-copy tokens often appeared in common boilerplate text (e.g. "stay at [NNP]" or "rates for [NNP]" followed by "Hotel").[15] Where we observe lower than expected rates (e.g. ") of [NNP]"), we find that the system may have memorized training sentences.

5.2 Analysis of Words That Are Not Copied

When words are not copied, what sort of output is the system producing? We find that it typically falls into one of four categories: *drop* (no target token aligns with the source token), *change* (the word is changed: partially translated, transliterated, or inflected even if it is not a target language word), *substitution* (the word is replaced with a fluent but not adequate substitute), or *translation* (translated into a target language word).

We begin with an automatic analysis. We randomly sample 200 examples each of sentences containing words that were not copied for novel-copy, novel-non-copy, observed-copy, and observed-non-copy NNPs (EN–DE). We retranslate each sentence and produce a soft alignment matrix from the attention mechanism, then convert the soft alignments between BPE segments into hard alignments between the source word and one or more target words.[16] A word has been *dropped* if it is unaligned. We count a word as being *changed* if any words it is aligned to have any subword (BPE segment) overlap with the original word's subwords. Both *substitution* and *translation* fall under *other*; we analyze those manually.

Results are shown in Table 3.[17] For all novel words, the most frequent output type is *change*. For example, the novel NNP *Bishnu* is changed into *Bischnu* in German.[18] Other changes include translations of parts of the word, and concatenation with other tokens. The output token often starts with the same character or sequence of characters as the source token.[19]

We manually inspect examples in the *other* category. For observed-non-copy words, almost all are translations (e.g. *Sea* translated correctly as *Meer*), as expected. For observed-copy words, we see a mix of translations and other changes to the words, which are almost evenly split between substitutions and small changes. These include inflections (e.g. *Bremen magazine* reasonably translated as *Bremer Magazin*[20]).

Within the *other* category, perhaps the most interesting cases are those where words appear to be substituted with a fluent but not adequate alternative. Many substitutions occur when the rare word is inserted next to a word that often forms a collocation (like "United States" – in sentences that include "in the [NNP] States" the translation sometimes defaults to a translation of "United States" regardless of the actual NNP inserted in place of "United"). Others have a less common NNP swapped for one that belongs to a similar semantic category (e.g. the place name *Dublin* being generated instead of the less common *Halle* – as Arthur et al. (2016) and others observed). For novel-copy words labeled as *other*, three quarters are substitutions and one quarter exhibit small changes. The reverse is true for novel-non-copy words: the majority exhibit small changes while almost thirty percent are substitutions.

5.3 Properties of Copied Words

Certain words exhibit properties that make them more likely to be copied, regardless of context. At first glance, it seems unintuitive that the rate of copying of novel-copy words and novel-non-copy words differs (Fig. 1) – the model has never observed any of these words, and they are being presented in identical contexts – why does it differentiate between them? Doing so indicates that the model has learned what makes a sequence of

[15] Since hidden representations contain whole sentence information, right side context may influence copying too.

[16] We use AmuNMT (Junczys-Dowmunt et al., 2016), producing slightly different output. See Appendix C for details.

[17] Rows do not sum to 200 because some words in our random sample were copied by the the AmuNMT decoder.

[18] A near-transliteration – the "sh"/"sch" transformation is seen in EN–DE cognates, e.g. "ship" and "Schiff".

[19] Appendix D contains examples of this and more.

[20] *Bremen* and *Bremer* are unique BPE segments, so the *change* heuristic could not be applied.

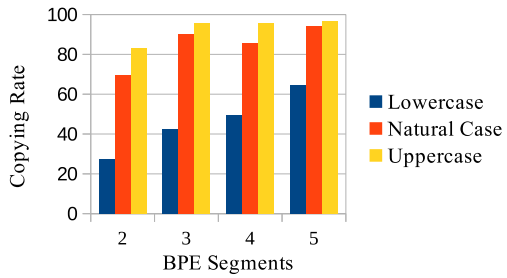

Figure 2: Copying rate based on casing and number of BPE segments for novel NNP words (EN–DE), averaged across all NNP contexts.

subwords likely to be copied.

Belinkov et al. (2017) observe that neural machine translation models may encode information about part-of-speech, which could be used when determining whether or not to copy (but does not explain within-POS differences). For numbers, it mainly learns to copy numerical portions while changing commas to periods and vice versa (as required by the target language's conventions). Nouns and proper nouns are more interesting: some should be translated (e.g. novel noun compounds like *hallmate*), or, in the case of misspellings (e.g. *manfacturer*), corrected, while others should be copied. For novel NN words, there is another striking difference between copy and non-copy: most of the former contain capital letters and most of the latter do not.

5.4 Capitalization and Copying

To experiment with the influence of capitalization on copying, we take each novel NNP word (96 copy and 22 non-copy) and convert it to lowercase, leave it in its natural case (all have at least one uppercase letter), or convert it to uppercase. We then translate all of them in all NNP contexts (from previous EN–DE experiments). Using only novel words sidesteps the issue of truecasing.

Lowercase words are the least frequently copied (average copy rate of 40.2%), uppercase words are the most copied (94.4%), and the natural case falls in the middle (81.7%). However, changing casing changes the BPE segmentation, and uppercase words tend to be split into more pieces: a mean of 4.4 segments, as compared to means of 3.1 (lowercase) and 2.9 (natural case). The number of subword segments correlates positively with copying rate (Fig. 2), but, controlling for that, we still find that NNP words that are completely capitalized tend to be copied more than those with the same number of subword segments but only lowercased letters, suggesting that the system is encoding information about the connection between capitalization and copying. We also perform this experiment with PROPN words in the DE–EN direction, and find that increased capitalization increases copying, though we do *not* find there that an increase in the number of BPE segments increases copying. The true casing of the word consistently falls between these two extremes. The high copying rate of fully-capitalized words is intuitive: acronyms are often both uppercased and copied from source to target. That is not to say that the model always learns to copy acronyms; it also learns to translate them when appropriate (such as *GDP* to *BIP*). There is always an interplay between learned translations and features that may encourage copying.

The connection between copying rate and capitalization provides one explanation for the gap in behavior of the two novel word types, and demonstrates that features of words influence copying. Note that it learns this behavior based on training data, without access to information at a finer granularity (character-level) than the subword units.

6 Conclusion

We show that subword vocabulary neural machine translation systems learn about copying from context and the subwords themselves. The effect of context is strong enough to cause words that would otherwise be translated to be copied. Characteristics of subword tokens play a role in copying behavior, with capitalized tokens more likely to be copied. We leave as future work a deeper analysis of the level of character-awareness encoded in representations of the BPE segments as a byproduct of training. We provide an analysis of what happens when words are not copied, showing expected differences between novel words and words that were observed during training. Additionally, we provide more examples and evidence of the problem of substituting fluent but non-adequate translations for rare or unknown words.

Acknowledgments We thank the reviewers and our colleagues for comments and suggestions. Rebecca Knowles was supported by a National Science Foundation Graduate Research Fellowship under Grant No. DGE-1232825. This work was also funded by the IARPA MATERIAL project.

References

Philip Arthur, Graham Neubig, and Satoshi Nakamura. 2016. Incorporating discrete translation lexicons into neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1557–1567, Austin, Texas. Association for Computational Linguistics.

Yonatan Belinkov, Lluís Màrquez, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. 2017. Evaluating layers of representation in neural machine translation on part-of-speech and semantic tagging tasks. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1–10, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Anna Currey, Antonio Valerio Miceli Barone, and Kenneth Heafield. 2017. Copied monolingual data improves low-resource neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 148–156, Copenhagen, Denmark. Association for Computational Linguistics.

Jiatao Gu, Zhengdong Lu, Hang Li, and Victor O.K. Li. 2016. Incorporating copying mechanism in sequence-to-sequence learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1631–1640, Berlin, Germany. Association for Computational Linguistics.

Caglar Gulcehre, Sungjin Ahn, Ramesh Nallapati, Bowen Zhou, and Yoshua Bengio. 2016. Pointing the unknown words. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 140–149, Berlin, Germany. Association for Computational Linguistics.

Marcin Junczys-Dowmunt, Tomasz Dwojak, and Hieu Hoang. 2016. Is neural machine translation ready for deployment? a case study on 30 translation directions. In *Proceedings of the 9th International Workshop on Spoken Language Translation (IWSLT)*, Seattle, WA.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, Andr F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in c++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Huda Khayrallah and Philipp Koehn. 2018. On the impact of various types of noise on neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 74–83, Melbourne, Australia. Association for Computational Linguistics.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Thang Luong, Ilya Sutskever, Quoc Le, Oriol Vinyals, and Wojciech Zaremba. 2015. Addressing the rare word problem in neural machine translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 11–19, Beijing, China. Association for Computational Linguistics.

Myle Ott, Michael Auli, David Grangier, and Marc'Aurelio Ranzato. 2018. Analyzing uncertainty in neural machine translation. In *International Conference on Machine Learning*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Edinburgh neural machine translation systems for WMT 16. In *Proceedings of the First Conference on Machine Translation (WMT)*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 173–180. Association for Computational Linguistics.

Appendices

A Collection of Left Bigram Contexts

We use left bigram contexts as a proxy to evaluate the contexts in which words are copied. Here, we provide details of the context selection process described in Section 5.1. In overview: for each POS, we collect the full set of left bigram contexts that ever precede a word with that tag, then filter by frequency and subsequent token diversity.

For each POS, first, we find all left bigram contexts $(tok0, tok1, copiedword)$ that occur in the training data (where the copied word was tagged with the given POS). We then filter this set so that it only contains contexts $(tok0, tok1)$ that appeared at least 1000 times (left of NNP/PROPN) or 500 times (left of CD/NN/NUM/NOUN) in the training data. To ensure that we're not simply capturing collocations ("European Union"), we filter out left bigram contexts that have been followed by fewer than 150 unique types with that particular POS. This results in between 53 and 276 contexts, as shown in Table 4.

POS	Num. Contexts
NNP	176
NN	82
CD	74
PROPN	276
NOUN	66
NUM	53

Table 4: Context counts by POS tag (NNP, NN, CD for EN–DE; PROPN, NOUN, NUM for DE–EN), selected as described in Appendix A.

Each context is then associated with a copying rate, calculated as the number of times the token (with the given POS tag) following $(tok0, tok1)$ is copied, divided by the total number of times $(tok0, tok1)$ was observed to be followed by a token with that POS tag. In Table 5, we show the most- and least-copy-prone contexts for EN–DE (those with the highest and lowest copying rates).

B Collection and Labeling of Copy/Non-Copy Words

In Section 5.1, we give a high-level description of how we collect and label words. All words that we examine are labeled as either *copy* or *non-copy*. For words that were observed in training, we filter out those that appeared fewer than 1000 times. We label them as *copy* if they were copied $\geq 70\%$ of the time in training data (according to the heuristic described in Section 4), and as *non-copy* if they were copied $\leq 30\%$ of the time in training data, discarding the remainder. For words that were unobserved in training, we used the same threshold but calculate it over all instances in the test data (with no requirement that they appear a certain number of times). Table 6 shows the number of words selected after filtering and thresholding.

POS	Context	Copy Rate
NNP	Finance Minister	94.5%
	rates for	94.0%
	congratulate Mr	91.7%
	between the	10.5%
	President ,	7.7%
CD	updated on	94.0%
	the B	0.1%
NN	notified when	97.3%
	the first	0.6%

Table 5: Left bigram contexts with the highest/lowest copying rates (EN–DE), by POS tag.

	Novel		Observed	
POS	**Copy**	**Non-C.**	**Copy**	**Non-C.**
NNP	96	22	251	263
NN	14	16	13	1664
CD	3	29	60	44
PROPN	92	76	463	418
NOUN	12	222	29	2176
NUM	2	29	55	68

Table 6: Counts of each word type by novel/observed, copy/non-copy distinction and POS tag (NNP, NN, CD are EN–DE; PROPN, NOUN, NUM are DE–EN).

C Attention and Alignments

We produce soft alignments (the attention matrix) using the AmuNMT decoder with the "return-nematus-alignment" flag set. It performs normalization differently than Marian's decoder (producing slightly different outputs for many sentences, including sometimes copying words that were not copied in our original translations).

For each target (subword) token, we align it to the source (subword) token with the highest

soft alignment weight. Given our source word of interest s (composed of subword segments $s_1 \dots s_n$), we define its translation to be the list of all target words t (composed of subword segments $t_1 \dots t_m$) for which any subword t_i was aligned to a subword s_j of s.

D Additional Examples

D.1 Changes

Here we show additional examples of *changes*, like the transliteration-like change of *Bishnu* to *Bischnu* described in Section 5.2.

There are also partial translations when BPE segments are full source language words – like *Thneed* (segmented "Th@@ need") becoming *ThNotwendigkeit* (segmented "Th@@ Notwendigkeit" – *Notwendigkeit* is a valid translation of *need*). Sometimes, a token *is* copied but then concatenated with another token.

Even without overlap of BPE segments between the source and the translation, changed words sometimes share a number of characters (especially at the beginning or end of a word). Half of the *other* category output of *Thneed* ("Th@@ need") begin with the letter "T" (but not the BPE token "Th@@"). This may suggest some level of character-awareness in the representations of BPE segments, produced as a byproduct of training. We leave a deeper analysis of this to future work.

D.2 NNP Substitutions

Here we provide additional examples of substitutions, as seen in Section 5.2. These findings provide additional support and nuance to the study of this phenomenon of neural machine translation system errors.

Many substitutions occur when the rare word is inserted next to a word that often forms a collocation (like "United States" or "European Union" or "Madam President"). For example, in a template where "in the [NNP]" is followed by "States", inserting the NNP *Accies* results in "in the *Accies* States" – which was then translated by the system as "in den Vereinigten Staaten" (gloss: "in the United States").

We also observe examples that may have to do with a combination of (in)frequency of tokens and the context. For example, we have the novel NNP *Sloveina* (perhaps a misspelling of *Slovenia*), which is often replaced with *Slowaken* (*Slovakia*) when translated to German. In another sentence, we find that "this year, *Angela* expects" is translated to "in diesem Jahr erwartet *Merkel*" despite *Merkel* appearing nowhere in the source text. The first and last names of German chancellor Angela Merkel appear frequently together in training data, and thus likely have sufficiently similar representations. We see other similar substitutions: *Mitt* for *Romney*, *US* for *Obama*, and *Thomas* for *Sarah*. Sometimes a specific name is replaced with a title, such as "your prime minister, *York*" being translated as "ihr Premierminister, *Herr Präsident*" (glossed as "your prime minister, *Mr. President*").

E Additional Plots

Here we include plots for DE–EN PROPN. Fig. 3 shows context experiments. It shows similar trends to Fig. 1, but with a greater gap between novel- copy/non-copy words. As noted, the DE–EN capitalization experiments (Fig. 4) show the same trends as EN–DE in terms of capitalization (despite the capitalization of all nouns in German), but not in terms of numbers of BPE segments.

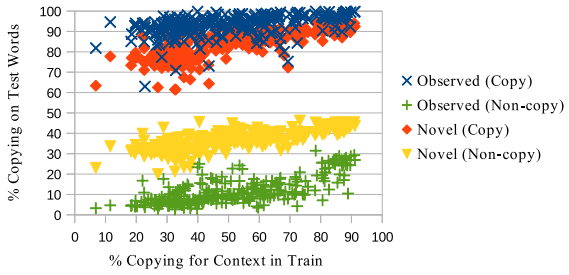

Figure 3: Percent of PROPN (DE–EN) tokens copied by how copy-prone the context is, by category. Each point is the percentage of copying for all within-category words, averaged across all example templates for one particular context.

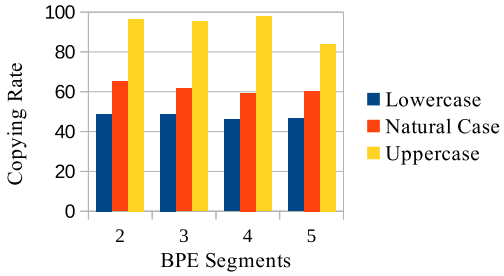

Figure 4: Copying rate based on casing and number of BPE segments for novel PROPN words (DE–EN), averaged across all PROPN contexts.